When is an outlier not an outlier?

Ken Rice
MRC Biostatistics Unit
5th April 2004
League Tables/Institutional Comparison

NHS to run death rate leagues

Whitehall cautious as Scots start table

or individual surgeons' mortality rates on the grounds that people would fail to appreciate the complexities involved, particularly differences in the health of patients on admission.

Death rates among patients having cail bladder removal
League Tables/Institutional Comparison

NHS to run death rate leagues

Schools in ‘worst borough’ improve

Whitehall cautious as Scots start table
League Tables/Institutional Comparison

NHS to run death rate leagues
Whitehall cautious as Scots start table

Schools in ‘worst borough’ improve

When is an outlier not an outlier?
Motivating example

Data from Commission for Health Improvement;

When is an outlier not an outlier?
Motivating example

Which rates are not equal to the average?
Motivating example

Which rates are not equal to the average?
Motivating example

Do a Bonferroni correction to allow for multiple comparisons
Motivating example

Do a Bonferroni correction to allow for multiple comparisons
Another example...
Another example...

- Bristol Royal Infirmary 1984-1995; heart operations on under 1’s;

- “...more children died than might have been expected...” [Public Inquiry]
Another example...

- Bristol Royal Infirmary 1984-1995; heart operations on under 1’s;
- “...more children died than might have been expected...” [Public Inquiry]
- Just ‘bottom of the league’, or more serious? How to quantify this?
Bristol Data

When is an outlier not an outlier?
Bristol Data

When is an outlier not an outlier?
Bristol Data

95%, 99% intervals assume known ‘null’ rate - note effect of big $n$
Analysing the Bristol Data fairly

- For most of the data, $Y_i \sim \text{Bin}(n_i, \theta), i = 1, \ldots, I$

- Outliers don’t follow this model, believe they are ‘more extreme’
Analysing the Bristol Data fairly

- For most of the data, $Y_i \sim \text{Bin}(n_i, \theta), i = 1, \ldots, I$

- Outliers don’t follow this model, believe they are ‘more extreme’

- Robust estimation; $M$-estimates, robust deviance methods, leads to robust $p$-values (more on this later!)
Analysing the Bristol Data fairly

- For most of the data, $Y_i \sim \text{Bin}(n_i, \theta), i = 1, \ldots, I$

- Outliers don’t follow this model, believe they are ‘more extreme’

- Robust estimation; $M$-estimates, robust deviance methods, leads to robust $p$-values (more on this later!)

- Details (all generous);
  - Robustly fit $\text{logit}(\theta) = \mu_0 + \beta \log(n_i)$, to allow for hospital size
  - Parametric bootstrap used on $p$-values, allows for parameter uncertainty
Analysing the Bristol Data fairly

- For most of the data, \( Y_i \sim \text{Bin}(n_i, \theta), i = 1, \ldots, I \)

- Outliers don’t follow this model, believe they are ‘more extreme’

- Robust estimation; \( M \)-estimates, robust deviance methods, leads to robust \( p \)-values (more on this later!)

- Details (all generous):
  - Robustly fit \( \text{logit}(\theta) = \mu_0 + \beta \log(n_i) \), to allow for hospital size
  - Parametric bootstrap used on \( p \)-values, allows for parameter uncertainty

<table>
<thead>
<tr>
<th>Method</th>
<th>Bristol’s ( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>raw MLE</td>
<td>0.0013</td>
</tr>
<tr>
<td>robust deviance</td>
<td>0.0008</td>
</tr>
<tr>
<td>( M )-estimate</td>
<td>0.0010</td>
</tr>
</tbody>
</table>
Multiple comparisons

- Still haven’t adjusted for the multiple comparisons problem
Multiple comparisons

- Still haven’t adjusted for the multiple comparisons problem
- New method for this: control False Discovery Rate (FDR);
Multiple comparisons

- Still haven’t adjusted for the multiple comparisons problem
- New method for this: control False Discovery Rate (FDR); See Benjamini and Hochberg, Storey, and others (genetics)
Multiple comparisons

- Still haven’t adjusted for the multiple comparisons problem

- New method for this: control False Discovery Rate (FDR);
  See Benjamini and Hochberg, Storey, and others (genetics)

- Controls \( \mathbb{E} (\text{proportion of ‘outliers’ found which are wrong}) \)
Multiple comparisons

- Still haven’t adjusted for the multiple comparisons problem
- New method for this: control False Discovery Rate (FDR); See Benjamini and Hochberg, Storey, and others (genetics)
- Controls $\mathbb{E}$ (proportion of ‘outliers’ found which are wrong)
- It’s easy!
Multiple comparisons

- Still haven’t adjusted for the multiple comparisons problem
- New method for this: control False Discovery Rate (FDR); See Benjamini and Hochberg, Storey, and others (genetics)
- Controls $\mathbb{E}$ (proportion of ‘outliers’ found which are wrong)
- It’s easy!
  - For ranked $p$-values $p_{(1)} \leq p_{(2)} \ldots \leq p_{(I)}$, 
- When is an outlier not an outlier?
Multiple comparisons

- Still haven’t adjusted for the multiple comparisons problem
- New method for this: control False Discovery Rate (FDR); See Benjamini and Hochberg, Storey, and others (genetics)
- Controls \( \mathbb{E}(\text{proportion of ‘outliers’ found which are wrong}) \)
- It’s easy!
  - For ranked \( p \)-values \( p(1) \leq p(2) \ldots \leq p(I) \), to control FDR at \( \alpha \), reject all data where \( p(i) \leq \frac{\alpha i}{I} \)
Multiple comparisons

- Still haven’t adjusted for the multiple comparisons problem

- New method for this: control False Discovery Rate (FDR); See Benjamini and Hochberg, Storey, and others (genetics)

- Controls $E$ (proportion of ‘outliers’ found which are wrong)

- It’s easy!
  - For ranked $p$-values $p(1) \leq p(2) \ldots \leq p(I)$, to control FDR at $\alpha$, reject all data where $p(i) \leq \frac{\alpha i}{I}$
  - So reject smallest $p$-value when $\leq \frac{\alpha}{I}$ (Bonferroni)
Multiple comparisons

- Still haven’t adjusted for the multiple comparisons problem
- New method for this: control False Discovery Rate (FDR); See Benjamini and Hochberg, Storey, and others (genetics)
- Controls \( \mathbb{E} \) (proportion of ‘outliers’ found which are wrong)
- It’s easy!
  - For ranked \( p \)-values \( p(1) \leq p(2) \ldots \leq p(I) \), to control FDR at \( \alpha \), reject all data where \( p(i) \leq \frac{\alpha i}{I} \)
  - So reject smallest \( p \)-value when \( \leq \frac{\alpha}{I} \) (Bonferroni)
  - Reject next smallest when \( \leq \frac{2\alpha}{I} \) (not Bonferroni)
Multiple comparisons

- Still haven’t adjusted for the multiple comparisons problem

- New method for this: control False Discovery Rate (FDR); See Benjamini and Hochberg, Storey, and others (genetics)

- Controls \( \mathbb{E} \) (proportion of ‘outliers’ found which are wrong)

- It’s easy!
  - For ranked \( p \)-values \( p(1) \leq p(2) \ldots \leq p(I) \),
    to control FDR at \( \alpha \), reject all data where \( p(i) \leq \frac{\alpha i}{I} \)
  - So reject smallest \( p \)-value when \( \leq \frac{\alpha}{I} \) (Bonferroni)
  - Reject next smallest when \( \leq 2\frac{\alpha}{I} \) (not Bonferroni)

- Define \( q_i \) as max FDR \( \alpha \) such that \( Y_i \) gets rejected; \( q(i) = \frac{p(i)I}{i} \)

- Intuitively; \( q_i = (\text{max}) \) FDR if \( Y_i \) and everything more extreme classed as outlier
Why is this useful here?

- Bonferroni controls $\mathbb{P}(\geq 1 \text{ ‘outlier’ found } | \text{ none actually there })$
Why is this useful here?

- Bonferroni controls $\mathbb{P}(\geq 1 \text{ ‘outlier’ found } | \text{ none actually there })$ - who cares?!
Why is this useful here?

- Bonferroni controls $\mathbb{P}(\geq 1 \text{ ‘outlier’ found | none actually there })$ - who cares?!

- FDR is of much more interest - how much embarrassment do we have to live with?
Why is this useful here?

- Bonferroni controls $\mathbb{P}(\geq 1 \text{ ‘outlier’ found } | \text{ none actually there })$ - who cares?!

- FDR is of much more interest - how much embarrassment do we have to live with?

- Realistically, have to exclude everyone beyond some limit, else unfair

- FDR makes sense of such behaviour
Why is this useful here?

- Bonferroni controls $\mathbb{P}(\geq 1 \text{ ‘outlier’ found | none actually there})$ - who cares?!

- FDR is of much more interest - how much embarrassment do we have to live with?

- Realistically, have to exclude everyone beyond some limit, else unfair

- FDR makes sense of such behaviour

- Bristol’s $q$-value lie between 0.010, 0.016

- Rejecting Bristol and everything more extreme, very low FDR
Bristol data via $q$ values

Bristol isn’t singled out in this analysis; if no outliers, $q_i$ always ‘unremarkable’
Return to motivating example

Still doesn’t stop us classing e.g. 70% as outliers here!
Asking the wrong question

- In this problem $Y_i \sim N(\mu_i, \sigma_i^2)$, $\sigma_i$ are known

- Ridiculous to test each $Y_i$ equal to some null value, $H_{0i} : \mu_i = \mu$
Asking the wrong question

- In this problem $Y_i \sim N(\mu_i, \sigma_i^2)$, $\sigma_i$ are known

- Ridiculous to test each $Y_i$ equal to some null value, $H_{0i} : \mu_i = \mu$

- Much better to test $H_{0i} : \mu_i \sim F_0$
Asking the wrong question

• In this problem $Y_i \sim N(\mu_i, \sigma_i^2)$, $\sigma_i$ are known

• Ridiculous to test each $Y_i$ equal to some null value, $H_{0i} : \mu_i = \mu$

• Much better to test $H_{0i} : \mu_i \sim F_0$

• Suggest that $F_0$ should be $N(\mu_0, \sigma_0^2)$
Asking the wrong question

- In this problem $Y_i \sim N(\mu_i, \sigma_i^2)$, $\sigma_i$ are known
- Ridiculous to test each $Y_i$ equal to some null value, $H_{0i}: \mu_i = \mu$
- Much better to test $H_{0i}: \mu_i \sim F_0$
- Suggest that $F_0$ should be $N(\mu_0, \sigma_0^2)$
- $\mu_0, \sigma_0$ are interpretable and usually of interest
- $N(\mu_0, \sigma_0^2)$ reasonable, and gives $Y_i \sim N(\mu_0, \sigma_i^2 + \sigma_0^2)$
Asking the wrong question

- In this problem $Y_i \sim N(\mu_i, \sigma_i^2)$, $\sigma_i$ are known

- Ridiculous to test each $Y_i$ equal to some null value, $H_{0i} : \mu_i = \mu$

- Much better to test $H_{0i} : \mu_i \sim F_0$

- Suggest that $F_0$ should be $N(\mu_0, \sigma_0^2)$

- $\mu_0, \sigma_0$ are interpretable and usually of interest

- $N(\mu_0, \sigma_0^2)$ reasonable, and gives $Y_i \sim N(\mu_0, \sigma_i^2 + \sigma_0^2)$

... for $Y_i$ which follow the null
Asking the wrong question

- In this problem \( Y_i \sim N(\mu_i, \sigma_i^2) \), \( \sigma_i \) are known

- Ridiculous to test each \( Y_i \) equal to some null value, \( H_{0i} : \mu_i = \mu \)

- Much better to test \( H_{0i} : \mu_i \sim F_0 \)

- Suggest that \( F_0 \) should be \( N(\mu_0, \sigma_0^2) \)

- \( \mu_0, \sigma_0 \) are interpretable and usually of interest

- \( N(\mu_0, \sigma_0^2) \) reasonable, and gives \( Y_i \sim N(\mu_0, \sigma_i^2 + \sigma_0^2) \)
  ... for \( Y_i \) which follow the null

- Need robust estimation for hierarchical models – still an open problem
More on $M$-estimation

Non-hierarchical: Instead of fitting $N(\mu_0, 1)$, fit $\mu_i \sim (1 - \epsilon)N(\mu_0, 1) + \epsilon F_1$;

When marginal is Normal on $[\mu_0 - k, \mu_0 + k]$ and exponential beyond, estimate is “most robust”. Huber (1964), other derivations available
More on $M$-estimation

Non-hierarchical: Instead of fitting $N(\mu_0, 1)$, fit $\mu_i \sim (1 - \epsilon)N(\mu_0, 1) + \epsilon F_1$;

When marginal is Normal on $[\mu_0 - k, \mu_0 + k]$ and exponential beyond, estimate is “most robust”. Huber (1964), other derivations available
More on $M$-estimation

Non-hierarchical: Instead of fitting $\mathcal{N}(\mu_0, 1)$, fit $\mu_i \sim (1 - \epsilon)\mathcal{N}(\mu_0, 1) + \epsilon F_1$;

When marginal is Normal on $[\mu_0 - k, \mu_0 + k]$ and exponential beyond, estimate is “most robust”. Huber (1964), other derivations available

We generalize; fit Normal on $[\mu_0 - k\sigma_0, \mu_0 + k\sigma_0]$, exponential beyond these limits. $\mu_0, k, \sigma_0$ all allowed to vary. Use $\mathcal{N}(\hat{\mu}_0, \hat{\sigma}_0^2)$ as $F_0$. 
Application to motivating example

Can fit either through MLE or WinBUGS (neither trivial);
Application to motivating example

Gives a meaningful measure of ‘outlying-ness’;

When is an outlier not an outlier?
Summary and Discussion

- Original application
  - CHI got 70% outliers because it asked the wrong question
Summary and Discussion

- Original application
  - CHI got 70% outliers because it asked the wrong question
  - Random effects models will be used from next year
Summary and Discussion

- Original application
  - CHI got 70% outliers because it asked the wrong question
  - Random effects models will be used from next year

- False Discovery Rates...
  - tell us about tail behaviour, not about individual data points
  - are therefore appropriate for institutional comparison
Summary and Discussion

- Original application
  - CHI got 70% outliers because it asked the wrong question
  - Random effects models will be used from next year

- False Discovery Rates...
  - tell us about tail behaviour, not about individual data points
  - are therefore appropriate for institutional comparison

- Further work
  - FDR tells us when a \( p \)-value/residual is significantly small/large
  - This is better than a QQ plot
Summary and Discussion

- Original application
  - CHI got 70% outliers because it asked the wrong question
  - Random effects models will be used from next year

- False Discovery Rates...
  - tell us about tail behaviour, not about individual data points
  - are therefore appropriate for institutional comparison

- Further work
  - FDR tells us when a $p$-value/residual is significantly small/large
  - This is better than a QQ plot
  - Dependency between $p_i$ not accounted for - but simulations suggest unimportant
  - Robust methods not ‘solved’, especially in hierarchical models
  - Robustly fit any model - how much of the data looks like an outlier?
References

- This work
  - Rice and Spiegelhalter, “A simple diagnostic plot connecting robust estimation, outlier detection, and false discovery rates”, American Statistician, submitted

- False Discovery Rate
  - Benjamini and Hochberg, “Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing”, JRSSB, 1995

- Robust estimation
What should we say about $F_1$?

- $(1 - \epsilon)F_0 + \epsilon F_1$; $F_1$ is the distribution of the contaminants
- Don’t want to say *much* about $F_1$...
What should we say about $F_1$?

- $(1 - \epsilon)F_0 + \epsilon F_1$; $F_1$ is the distribution of the contaminants
- Don’t want to say *much* about $F_1$...

- Cauchy(1) can be 60% $N(0,2)$ and 40% $F_1$ - or lots of other mixtures
What should we say about $F_1$?

- $(1 - \epsilon)F_0 + \epsilon F_1$; $F_1$ is the distribution of the contaminants

- Don’t want to say much about $F_1$...

- Cauchy(1) can be 60% $N(0,2)$ and 40% $F_1$ - or lots of other mixtures

- Have to say something about $F_1$ to get anywhere at all